Naive possibilistic classifiers for imprecise or uncertain numerical data

Myriam Bounhas, Mohammad Ghasemi Hamed, Henri Prade, Mathieu Serrurier, Khaled Mellouli

Naive Possibilistic Classifiers for Imprecise or Uncertain Numerical Data

Myriam Bounhas[a,b,*], Mohammad Ghasemi Hamed[c,d], Henri Prade[c], Mathieu Serrurier[c], Khaled Mellouli[a]

[a] *LARODEC Laboratory, ISG de Tunis, 41 rue de la Liberté, 2000 Le Bardo, Tunisie*
[b] *Emirates College of Technology, P.O. Box: 41009, Abu Dhabi, United Arab Emirates*
[c] *IRIT, UPS-CNRS, 118 route de Narbonne, 31062 Toulouse Cedex09, France*
[d] *DTI-R&D DGAC - 7 avenue Edouard Belin 31055 Toulouse, France*

Abstract

In real-world problems, input data may be pervaded with uncertainty. In this paper, we investigate the behavior of naive possibilistic classifiers, as a counterpart to naive Bayesian ones, for dealing with classification tasks in presence of uncertainty. For this purpose, we extend possibilistic classifiers, which have been recently adapted to numerical data, in order to cope with uncertainty in data representation. Here the possibility distributions that are used are supposed to encode the family of Gaussian probabilistic distributions that are compatible with the considered data set. We consider two types of uncertainty: i) the uncertainty associated with the class in the training set, which is modeled by a possibility distribution over class labels, and ii) the imprecision pervading attribute values in the testing set represented under the form of intervals for continuous data. Moreover, the approach takes into account the uncertainty about the estimation of the Gaussian distribution parameters due to the limited amount of data available. We first adapt the possibilistic classification model, previously proposed for the certain case, in order to accommodate the uncertainty about class labels. Then, we propose an algorithm based on the extension principle to deal with imprecise attribute values. The experiments reported show the interest of possibilistic classifiers for handling uncertainty in data. In particular, the probability-to-possibility transform-based classifier shows a robust behavior when dealing with imperfect data.

Keywords: Naive possibilistic classifier, possibility theory, numerical data, Naive Bayesian classifier, uncertainty

*Corresponding author
Email addresses: Myriam_Bounhas@yahoo.fr (Myriam Bounhas), mohammad.gh@gmail.com (Mohammad Ghasemi Hamed), prade@irit.fr (Henri Prade), Serrurier@irit.fr (Mathieu Serrurier), Khaled.Mellouli@topnet.tn (Khaled Mellouli)

1. Introduction

Uncertain or imprecise data may be encountered for instance when sensors are not fully reliable or when experts are not fully sure about the right class. Since standard classification techniques are not suitable to cope with imperfection in data, a common way to deal with this situation is to ignore such data. This then leads to a loss of information and the models obtained are not a faithful representation of reality.

Possibility theory [27] [23] has been recently proposed as a counterpart of probability theory to deal with classification tasks in presence of uncertainty. The study of naive possibilistic classifiers is motivated by the simplicity and the acceptable performances of Naive Bayesian Classifiers (NBC for short) and by the ability of possibility theory [27] to handle poor data. In spite of the fact that possibility distributions are useful for representing imperfect knowledge, there have been only few works that use possibility theory for classification [5] and most of existing Naive Possibilistic Classifiers deal with categorical attributes only.

For this reason, we study Naive Possibilistic Classifiers (NPC for short) that are based on the possibilistic counterpart of the Bayesian formula [28] and the estimation of possibility distributions from numerical data. Given a new piece of data to classify, a possibilistic classifier estimates its plausibility of belonging to each class (according to the training set of examples), and assigns the class having the highest plausibility value. In order to do that, we have to build a possibility distribution from data. Even if we assume that the data follow a Gaussian probabilistic distribution, its parameters are estimated from a limited sample set, and are then necessarily pervaded with imprecision. In this scope, we consider the possibility distribution that encodes all the Gaussian distributions for which the parameters are in a chosen confidence interval. Then, we extend the method in order to cope with the uncertainty in data sets. In other words, two types of uncertainty are taken into account: the uncertainty in the description of the data (both for the classes in the training set and the attribute values in the testing set), and the uncertainty due to the limited amount of data.

This work is a fully revised and an extended version of a conference paper [12]. While this latter paper extends the possibilistic classifiers proposed in [10] to handle uncertainty in data, this work also investigates a new way of building a possibility distribution as the representation of a family of probability distributions. This allows us to define two new possibilistic classifiers, called here NPC-2 and FNPC-2 for a flexible counterpart of the previous one. Moreover, the experimental part relies on a larger number of benchmarks with perfect as well as imperfect data.

The paper is structured as follows. Section 2 reviews some related works. In Section 3, we restate and motivate the idea of a possibilistic classifier. Section 4 introduces the method for computing possibilistic distributions that upper bound a family of Gaussian distributions. In Section 5, we extend possibilistic classifiers to the handling of uncertainty in the description of data: first for

the processing of *uncertain* classes in the training set, and then for dealing with imprecise attribute values modeled by intervals, in the testing set. The experimentation results are given in Section 6. The experiments reported show the interest of possibilistic classifiers to deal with perfect and imperfect data. Finally, Section 7 concludes and suggests some directions for future research.

2. Related Works

Some approaches have already proposed the use of a possibilistic data representation in classification methods that are based on decision trees, Bayesian-like, or case-based approaches. A general discussion about the appropriateness of fuzzy set methods in machine learning can be found in [39]. Most of the works in possibilistic classification are motivated by the handling of imprecision and uncertainty about attribute values or classes. Some assume that there is a partial ignorance about class values. This ignorance, modeled through possibility distributions, reflects the expert knowledge about the possible class of each training instance.

In general, the approaches deal with discrete attribute values only and are not appropriate for numerical attributes (and thus require a preliminary discretization phase for handling such attribute values). By contrast, the work reported here presents a new type of classification method suitable for classifying data pervaded with uncertainty. Our approach can handle numerical attributes, which can be imprecise in the testing set, and uncertain classes. It also takes into account the amount of data available. All these forms of uncertainty are represented in the possibility theory setting.

We now provide a brief survey of the literature on possibilistic classification. We start with approaches based on decision trees, before a more detailed discussion on Bayesian classifiers applied to possibilistic data.

Denoeux and Zouhal [55] use possibility theory to model and deal with uncertain labels in the training set. To do this, the authors assign a possibility degree to each possible class label which reflects the possibility that the given instance belongs to this class. Besides, Ben Amor et al.[3] have developed a qualitative approach based on decision trees for classifying examples having uncertain attribute values. Uncertainty on attribute values is represented by means of possibility distributions given by an expert. In [40], possibilistic decision trees are induced from instances associated with categorical attributes and vaguely specified classes. Uncertainty, modeled through possibility theory, concerns only the class attribute whereas other predictive attributes are supposed to be certainly known. The authors developed three approaches for possibilistic decision trees. The first one, using possibility distributions at each step of the tree construction, is based on a measure of non-specificity in possibility theory in order to define an attribute selection measure. The two remaining approaches make use of the notion of similarity between possibility distributions for extending the C4.5 algorithm in order to support data uncertainty.

A naive Bayesian-like possibilistic classifier has been proposed by Borgelt et al.[8] to deal with imprecise training sets. For this classifier, imprecision concerns only attribute values of instances (the class attribute and the testing set are supposed to be perfect). Given the class attribute, possibility distributions for attributes are estimated from the computation of the maximum-based projection [9] over the set of precise instances which contains both the target value of the considered attribute and the class.

A naive possibilistic network classifier proposed by Haouari et al. [37], presents a procedure that deals with training datasets with imperfect attributes and classes, and a procedure for classifying unseen examples which may have imperfect attribute values. This imperfection is modeled through a possibility distribution given by an expert who expresses his partial ignorance, due to a lack of prior knowledge. There are some similarities between our proposed approach and the one by [37]. In particular, they are both based on the idea stating that an attribute value is all the more possible if there is an example, in the training set, with the same attribute value in the discrete case, or a very close attribute value in terms of similarity in the numerical case. However, the approach in [37] does not require any conditional distribution over attributes to be defined in the certain case, whereas a preliminary requirement in our approach, is to estimate such a possibility distribution for numerical data in the certain case.

Benferhat and Tabia [5] propose an efficient algorithm for revising, using Jeffrey's rule, possibilistic knowledge encoded by a naive product-based possibilistic network classifier on the basis of uncertain inputs. The main advantage of the proposed algorithm is its capability to process the classification task in polynomial time with respect to the number of attributes.

In [51], the authors propose a new Bayesian classifier for uncertain categorical or continuous data by integrating uncertainty in the Bayesian theorem and propose a new parameter estimation method. An attempt to treat uncertainty in continuous data is proposed in [52], where authors developed a classification algorithm able to generate rules from uncertain continuous data. For the two works [52], [51], uncertainty over continuous attribute values is represented by means of intervals. This imprecision is handled by a regular probabilistic approach.

Besides, some case-based classification techniques, which make use of possibility theory and fuzzy sets, are also proposed in the literature. We can particularly mention the possibilistic instance-based learning approach [38]. In this work, the author proposes a possibilistic version of the classical instance-based learning paradigm using similarity measures. Interestingly, this approach supports classification and function approximation at the same time. Indeed, the method is based on a general possibilistic extrapolation principle that amounts to state that the more similar to a known example the case to be classified is, the more plausible the case and the example belong to the same class. This idea is further refined in [38] by evaluating this plausibility by means of an interval whose lower bound reflects the "guaranteed" possibility of the class, and the upper bound the extent to which this class is not impossible. In a more recent work [6], the authors develop a bipolar possibilistic method for case-based

learning and prediction.

This possibilistic instance-based learning approach may look similar to the proximity-based classifiers proposed in [10]. However, there are differences, although both emphasize a possibilistic view of classification based on similarity. In [38] a conditional possibility of a class given the case description is defined directly, taking into account all the attributes together. In the methods presented in [10], we rather start by defining the plausibility of a particular attribute value for a given class (on a similarity basis), and then apply a Bayesian-like machinery for obtaining the classification result.

3. General setting of possibilistic classification

We first recall some basics of possibility theory and then present the possibilistic classification viewed as a possibilistic version of the Bayes rule. In the following we also motivate the potential interest of possibility theory in classification.

3.1. Basic notions of possibility theory

Possibility theory [57, 23, 27] handles epistemic uncertainty in a qualitative or quantitative way. In particular, possibility theory is suitable for the representation of imprecise information.

Possibility theory is based on *possibility distributions*. Given a universe of discourse $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, a possibility distribution π is a function that associates to each element ω_i in the universe of discourse Ω a value in a bounded and linearly ordered valuation set $(L, <)$. This value is called a *possibility degree*. This scale may be quantitative, or qualitative when only the ordering between the degrees makes sense. In this paper, possibility degrees have a *quantitative* reading and L is taken as the unit interval [0,1]. A possibility distribution is used as an elastic constraint that restricts the more or less possible values of a single-valued variable.

Possibility distributions have two types of quantitative interpretations. The first one, that is related to fuzzy set theory, is the representation of gradual properties. For instance, linguistic expressions such that "long", "old" or "expensive" do not refer to a specific value, but to a set of possible values in a given context. For instance, a possibility distribution over prices may reflect the information that an house in a particular area is "expensive". In such a case, each price will be associated with a possibility degree which quantifies how much this price is typical with respect to the concept "expensive". When assigned to events, possibility degrees represent the *plausibility* that the events will occur. However, a possibility distribution may then also be viewed as representing a family of probability distributions, as we are going to see (see [21] for an introduction to this use).

By convention, $\pi(\omega_i) = 1$ means that it is fully possible that ω_i is the value of the variable. Note that distinct values ω_i, ω_j may be such that $\pi(\omega_i) = 1 = \pi(\omega_j)$. Besides, $\pi(\omega_i) = 0$ means that ω_i is impossible as the value of

the variable. Thanks to the use of the interval [0,1], intermediary degrees of possibility can be assessed, which enables us to acknowledge that some values are more possible than others.

In possibility theory, different important particular cases of knowledge situation can be represented:

- Complete knowledge: $\exists \omega_i, \pi(\omega_i) = 1$ *and* $\forall \omega_i \neq \omega_j, \pi(\omega_j) = 0$.
- Partial ignorance: $\forall \omega_i \in A \subseteq \Omega, \pi(\omega_i) = 1$, $\forall \omega_i \notin A, \pi(\omega_i) = 0$ (when A is not a singleton).
- Total ignorance: $\forall \omega_i \in \Omega, \pi(\omega_i) = 1$ (*all values in* Ω *are possible*).

A possibility distribution π on Ω enables events to be qualified in terms of their plausibility and their certainty, by means of two dual possibility and necessity measures that are respectively defined for an event $A \subseteq 2^{\Omega}$ by the formulas:

$$\Pi(A) = max_{\omega \in A} \pi(\omega)$$

$$N(A) = min_{\omega \notin A}(1 - \pi(\omega)) = 1 - \Pi(\overline{A})$$

The possibility measure $\Pi(A)$ evaluates to which extent A is consistent with our knowledge represented by π. Indeed, the evaluation provided by $\Pi(A)$ corresponds to a degree of *non-emptiness of the intersection* of the classical subset A with the fuzzy set having π as membership function. Moreover, $N(A)$ evaluates to which extent A is certainly implied by our knowledge, since it is a degree of *inclusion* of the fuzzy set corresponding to π into the subset A.

Quantitative possibility distributions can represent, or more generally approximate, a family of probability measures [25]. Indeed, a possibility measure Π can be viewed as an upper bound of a probability measure, and associated with the family of probability measures defined by

$$\mathcal{P}(\Pi) = \{P \text{ s. t. } \forall A, \Pi(A) \geq P(A)\}.$$

Thanks to the duality between Π and N and the auto-duality of P ($P(A) = 1 - P(\overline{A})$), it is clear that

$$\forall P \in \mathcal{P}(\Pi), \forall A, \Pi(A) \geq P(A) \geq N(A).$$

This is the starting point for defining a probability-possibility transform. The width of the gap between $N(A)$ and $\Pi(A)$ evaluates the amount of ignorance about $P(A)$, since it corresponds to the interval containing the imprecisely known probability. Thus, possibility distributions can in particular represent precise or imprecise information (representable by classical subsets) as well as complete ignorance. The possibilistic representation of complete ignorance should not be confused with a uniform probability distribution. Indeed, with

the above representation, we have $\Pi(A) = 1$ for any non empty event A, and $N(A) = 0$ for any event A different from Ω, while a uniform probability distribution on a universe with more than two elements associates non trivial events with a probability degree strictly between 0 and 1, which sounds paradoxical for a situation of complete ignorance. Possibility theory is particularly suited for representing situations of partial or complete ignorance (see [21], [29] for detailed comparative discussions between probability and possibility).

3.2. Conditional Possibility and Possibilistic Bayesian Rule

Conditioning in possibility theory is defined through a counterpart of Bayes rule, namely

$$\Pi(A \cap B) = \Pi(A|B) * \Pi(B)$$

It is has been shown that there are only two basic choices for $*$, either minimum or the product [24]. The min operator is suitable in the qualitative possibility theory setting, while the product should be used in quantitative possibility theory [16]. Quantitative possibilistic conditioning can be viewed as a particular case of Dempster's rule of conditioning since possibility measures are special cases of plausibility functions [54].

Thus, possibilistic conditioning corresponds to revising an initial possibility distribution π, when a new information $B \subseteq \Omega$ becomes available. In the quantitative setting we have:

$$\pi(a \mid B) = \begin{cases} \frac{\pi(a)}{\Pi(B)} & \text{if } a \in B \\ 0 & \textit{otherwise.} \end{cases}$$

3.3. Naive Bayesian Possibilistic Classification

The idea of applying possibility theory to classification parallels the use of probabilities in Bayesian classifiers (see the Appendix for a reminder). Probability distributions used in NBCs are usually built by assuming that numerical attributes are normally distributed around their mean. Even if a normal distribution is appropriate, identifying it exactly from a sample of data is especially questionable when data are poor. Gaussian kernels can be used for approximating any type of distributions which sounds more reasonable when normality assumptions are violated. Then, it is required to assess many parameters, a task that may be not compatible with poor data. The problem of the precise estimation of probability distributions for NBCs is important for the exact computation of the probability distribution over the classes. However, due to the use of the product for combining probability values (which are often small), the errors on probability estimations may have a significant effect on the final estimation. This contrasts with possibility distributions which are less sensitive to imprecise estimation for several reasons. Indeed, a possibility distribution may be viewed as representing a family of probability distributions corresponding to imprecise probabilities, which is more faithful in case of poor data. Moreover,

we no longer need to assume a particular shape of probability distribution in this possibilistic approximation process.

In the spirit of Bayesian classification, possibilistic classification is based on the possibilistic version of the Bayes theorem. Given a new vector $\{a_1, ..., a_M\}$ of n observed variables $A_1, ..., A_M$ and the set of classes $C = \{c_1, ..., c_C\}$, the classification problem consists in estimating a possibility distribution on classes and in choosing the class with the highest possibility for the vector X in this quantitative setting, i.e.:

$$\pi(c_j|a_1, ..., a_M) = \frac{\pi(c_j) * \pi(a_1, ..., a_M|c_j)}{\pi(a_1, ..., a_M)} \tag{1}$$

In formula (1), the quantitative component of possibilistic classification involves a *prior* possibility distribution over the classes and a *prior* possibility distribution associated with the input variables. Note that the term $\pi(a_1, ..., a_M)$ is a normalization factor and it is the same over all class values. In case we assume that there is no a priori knowledge about classes and the input vector to classify, we have $\pi(c_j) = 1$ and $\pi(a_1, ..., a_M) = 1$. Moreover, analogously to naive Bayesian classification, naive possibilistic classification makes an independence hypothesis about the variables A_i in the context of classes [4].

Assuming attribute independence, the plausibility of each class for a given instance is computed as:

$$\pi(c_j|a_1, ..., a_M) = \frac{\pi(c_j) * \prod_{i=1}^{M} \pi(a_i|c_j)}{\pi(a_1, ..., a_M)} \tag{2}$$

where conditional possibilities $\Pi(a_i|c_j)$ in formula (2) represent to which extent a_i is a possible value for the attribute A_i in the presence of the class c_j. As in the case of the conditioning rule, $*$ (and $\prod$ by extension) may be chosen as the min or the product operator (min corresponds to complete logical independence, while the use of the product makes partially possible values jointly less possible). Note that if we assume that there is no prior knowledge about classes, the term $\pi(c_j)$ can be omitted. In a product-based setting, a given instance is assigned to the most plausible class c^*:

$$c^* = arg \max_{c_j} (\pi(c_j) * \prod_{i=1}^{M} \Pi(a_i|c_j)) \tag{3}$$

It is worth noticing that formula (2) has a set-theoretical reading. Namely, when the possibility distributions take only the values 0 and 1, the formula (2) amounts to express that an instance may be possibly classified in c_j inasmuch as the attribute values of this instance are compatible with this class given the available observations. Thus, possibilistic classification may be viewed as an intermediate between Bayesian probabilistic classification and a purely set-based classifier (such classifiers use, for each attribute, the convex hull of the data values as a possibility distribution for identifying classes, usually leading to too many multiple classifications).

4. Computing a possibility distribution as a family of Gaussian distribution from a data sample

In this section, we explain how to build a possibility distribution from a set of data. First, we suppose that the data follow a Gaussian distribution with unknown parameters. By taking into account the uncertainty attached to the estimation of these parameters from a sample set, we propose to build the possibility distribution that encodes all the Gaussian distributions that may have generated the data with a chosen confidence level. Then, we extend this approach to Gaussian kernels.

4.1. Probability-possibility transformation

There are several transformations for moving from the probability framework to the possibility framework based on various principles such as consistency (what is probable is possible) or information invariance [14, 17, 43, 26, 22]. Dubois et al. [30] suggest to use the "maximum specificity" principle which aims at finding the most informative possibility distribution that encodes the considered probability information. A possibility distribution π_1 is more specific than a possibility distribution π_2 if and only if

$$\forall x \in \Omega, \pi_1(x) \leq \pi_2(x).$$

Since a possibility distribution explicitly handles the imprecision and is also based on an ordinal structure rather than an additive one, it has a weaker representation power than a probability one. This kind of transformation (probability to possibility) may be desirable when we are in presence of poor source of information, or when it is computationally harder to work with the probability measure than with the possibility measure.

In the case where the universe of discourse is discrete (i.e. $\Omega = \{c_1, \ldots, c_q\}$), the most specific possibility distribution π^* given a probability distribution p over Ω is defined by:

$$\forall i \in \{1, \ldots, q\}, \pi^*(c_i) = \sum_{c_j | p(c_j) \leq p(c_i)}^{q} p(c_j). \tag{4}$$

Example: For instance, we consider $\Omega = \{c_1, c_2, c_3\}$ and p such that $p(c_1) = 0.5$, $p(c_2) = 0.3$ and $p(c_3) = 0.2$. We obtain $\pi^*(c_1) = 0.5 + 0.3 + 0.2 = 1$, $\pi^*(c_2) = 0.3 + 0.2 = 0.5$ and $\pi^*(c_3) = 0.2$.

When the universe is continuous (i.e. $\Omega = \mathbb{R}$), the most specific possibility distribution function for a unimodal probability distribution function is given by [22]:

$$\pi_*(x) = sup\{1 - P(I^*_\beta), x \in I^*_\beta\} \tag{5}$$

where π_* is the most specific possibility distribution given a probability distribution p, I^*_β is the β-confidence interval ($P(I^*_\beta) = \beta$). Thus, if p has a finite

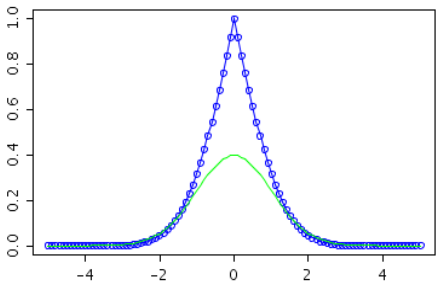

Figure 1: The maximum specific possibility distribution for $\mathcal{N}(0,1)$.

number of modes, π_* is the possibility distribution for which each $(1-\beta)$-cuts correspond to the confidence interval I_β^* of p. When p is unimodal, the unique value x such that $\pi_*(x) = 1$ is the mode of p. This possibility distribution is the most specific one whose associated possibility measure provides an upper bound for the considered probability measure.

Figure 1 presents the maximally specific probability-possibility transformation (in blue) of a normal distribution (in green).

4.2. Confidence region of the normal distribution parameters

Suppose that we have n observations $X_1, X_2, \cdots X_n$ drawn from a normal distribution with unknown mean μ and unknown variance σ^2. The $1-\alpha$ confidence region for the parameters of $\mathcal{N}(\mu, \sigma^2)$, contains a region in the two dimensional space of μ and σ^2 which has a probability equal to $1-\alpha$ to contain the true values of the parameters μ and σ^2. Arnold and Shavelle in [2] have compared several methods for finding such confidence regions. In their paper, they present the method that we describe below and they call it the Mood's method. The idea of Mood confidence region is to take α_1 and α_2 such as $1-\alpha = (1-\alpha_1)(1-\alpha_2)$, where $1-\alpha$ is the confidence level of the found region. Considering $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ and $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ respectively as the mean and the standard deviation estimated on the sample set, the confidence region $\mathcal{R}(n, \overline{X}, S)$ is defined by:

$$\mathcal{R}(n, \overline{X}, S) = \{(\mu, \sigma^2) : \sigma^2_{min} \leq \sigma^2 \leq \sigma^2_{min}, \mu_{min} \leq \mu \leq \mu_{max}\} \tag{6}$$

with

$$\sigma^2_{min} = \frac{n-1}{\chi^2_{1-\frac{\alpha_2}{2},n-1}} S^2$$
$$\sigma^2_{max} = \frac{n-1}{\chi^2_{\frac{\alpha_2}{2},n-1}} S^2$$
$$\mu_{min} = \overline{X} - \Phi_{1-\frac{\alpha_1}{2}} \frac{\sigma}{\sqrt{n}}$$
$$\mu_{max} = \overline{X} + \Phi_{1-\frac{\alpha_1}{2}} \frac{\sigma}{\sqrt{n}}.$$

where Φ_q and $\chi_{q,k}$ are respectively the q^{th} quantile of the standard normal distribution and the q^{th} quantile of the chi square distribution with k degrees of freedom. The authors also provide a table that indicates the optimal combination of α_1 and α_2 that gives the smallest possible region for a fixed confidence level $1-\alpha$ and for a fixed number of observations n.

By using the equation (6) we can find the mean and variance confidence interval respectively $[\mu_{min}, \mu_{max}]$,$[\sigma^2_{min}, \sigma^2_{max}]$, associated with our confidence region. Once we have found the confidence region, we define Θ as the family which contains all the probability functions p in the confidence region i.e.

$$\Theta = \{p = \mathcal{N}(\mu, \sigma^2) | (\mu, \sigma^2) \in \mathcal{R}(n, \overline{X}, S)\}.$$

4.3. Possibility distribution for a family of Gaussian distributions

We have shown how to build a confidence region for the parameters of a normal distribution (for a simplification purpose, we always take $1-\alpha = 0.95$ for the regions in the following). Since the estimation of these parameters is a critical issue for the naive Bayes classifier, it may be interesting to take into account the uncertainty around the parameters of the normal distribution that may have generated the data. In this scope, we propose to construct the most specific possibility distribution that contains the family Θ of Gaussian distributions that have mean and variance parameters in the confidence region.
We name $\Lambda = \{\pi | \pi = Tr(p), p \in \Theta\}$ the set of possibility distributions obtained by transforming each distribution in Θ ($Tr(p)$ is the possibility transform of a probability distribution using Formula 5). Thus, the possibility distribution defined by

$$\pi_{(n,\overline{X},S)}(x) = sup\{\pi(x) | \pi \in \Lambda\}$$

encodes all the family Θ. $\pi_{(n,\overline{X},S)}$ has the following definition:

$$\pi_{(n,\overline{X},S)}(x) = \begin{cases} 1 \text{ if } x \in [\mu_{min}, \mu_{max}] \\ 2 * \mathcal{G}(x, \mu_{min}, \sigma^2_{max}) & \text{if } x < \mu_{min} \\ 2 * \mathcal{G}(2 * \mu_{max} - x, \mu_{max}, \sigma^2_{max}) & x > \mu_{max} \end{cases} \quad (7)$$

where μ_{min}, μ_{max} and σ^2_{max} are respectively the lower and the upper bounds of the mean confidence interval, and the upper bound of the variance confidence

interval associated to the confidence region found by (6). Moreover, $\mathcal{G}(x, \mu, \sigma^2)$ is the cumulated distribution function of the $\mathcal{N}(\mu, \sigma^2)$. Possibility distributions encoding a family of probability distributions have been successfully applied to regression [35] where possibilistic $k-$NN regression consists in predicting intervals rather than precise values. For a detailed discussion about encoding probability distributions by possibility distributions, see [1, 36].

Figure 2 presents the distribution $\pi_{(10,\overline{X},S)}$ for the family of Gaussian distributions (in green) that are in the Mood region obtained from a sample of 10 pieces of data that follows the distribution $\mathcal{N}(0, 1)$.

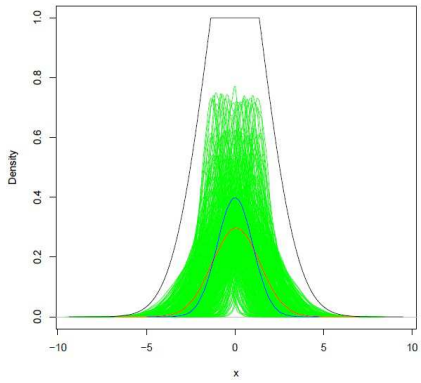

Figure 2: An example of the possibility distribution obtained for the family Θ, with a confidence level of 0.95 and a dataset with n=10.

4.4. Probability to possibility transformation-based classifiers

We apply the method presented above to naive Bayesian classifiers, where the distributions are assumed to be normal, and then to its flexible extension FNBC (using a combination of normal distributions). We shall call NPC-2 and FNPC-2 the possibilistic extensions of NBC and FNBC. In the possibilistic setting, we still assume that the probability distributions we start with are normal (or a combination of normal distributions), but we also encode the uncertainty around the estimations of their parameters.

In order to build NPC-2, we need to compute three types of possibility degrees: $\pi(c_i)$ the possibility of a class c_i, $\pi(a_i)$ the possibility of the attribute value a_i, and $\pi(a_i|c_j)$ the conditional possibility of a_i knowing c_j. These values are obtained as follows:

- $\pi(c_i)$ is obtained by computing the probability-possibility transformation (using equation 4) of the prior probability distribution over the classes;
- $\pi(a_i)$ is obtained by computing (eq. 7) the possibility distribution $\pi_{(N,\overline{X}_i,S_i)}$ that encodes the confidence region $\mathcal{R}_i(N, \overline{X}_i, S_i)$ for the parameters of

the normal distributions of A_i where N is the number of examples in the database, $\overline{X}_i$ is the means of the a_i values and S_i their standard deviation;

- $\pi(a_i|c_j)$ is obtained by computing (using equation 7) the possibility distribution $\pi_{(N_j,\overline{X}_{(i,j)},S_{(i,j)})}$ that encodes the confidence region for the parameters of the normal distributions of A_i (i.e. $\mathcal{R}_{(i,j)}(N_j, \overline{X}_{(i,j)}, S_{(i,j)})$) where N_j is the number of examples in the database that are associated to the class c_j, $\overline{X}_{(i,j)}$ is the means of the a_i values on this subset and $S_{(i,j)}$ their standard deviation.

The FNPC-2 is exactly the same as the NPC-2 in all respects, except that the method used for density estimation on continuous attributes is different. Rather than using a single Gaussian distribution for estimating each continuous attribute, we use a kernel density estimation as in FNBC. Kernel estimation with Gaussian kernels looks much the same except that the estimated density is averaged over a large set of kernels. For the FNPC-2, we use the following expression:

$$\pi(a|c_j) = \frac{1}{N_j} \sum_{k=1}^{N_j} \pi_{(N_j,\mu_{ik},\sigma_j)}(a) \tag{8}$$

where a is a value for the attribute A_i, k ranges over the N_j instances of the training set in class c_j and $\mu_{ik} = a_{ik}$ (a_{ik} is the value of the attribute A_i for the k-th example in the considered subset). For all distributions, the standard deviation is estimated by

$$\sigma_j = \frac{1}{\sqrt{N_j}}.$$

Besides, in this approach and for all the rest of this work, all attribute values a_i's are normalized using

$$a_{i_n} = \frac{a_i - min(a_i)}{max(a_i) - min(a_i)}.$$

5. Possibilistic distributions for imperfect numerical data

In many domains, databases are supplied with various information sources which may be neither fully reliable nor precise. That is why, available information is often pervaded with uncertainty and imprecision. In particular for numerical data, many factors contribute to make them imperfect, such as the variability of data, the use of unreliable data transmission or outdated sources, or the measurement errors. For instance, data provided by sensor networks such as temperature, pressure and rain measurement may be uncertain or imprecise.

In the context of classification or diagnosis problems, attributes in the training or testing sets may have uncertain numerical values. For instance, when classifying unseen examples, only an interval for a numerical value may be given instead of a precise value in some situations.

Imperfection can also affect categorical data and especially the training set labels. In fact, the framework of supervised learning which assumes precision and full certainty does not necessarily correspond to practical situations. The acquisition of a large volume of certain and precise labeled data may be problematic in some real domains, for cost reasons or partial lack of knowledge on the problem to be solved. It is often the case when the data are labeled by experts. An expert may be more comfortable in expressing uncertainty or imprecision in this task. Indeed the precise qualification of a situation by an expert may be difficult (e.g., in medical diagnosis, or in law application data).

Since standard classification techniques are inappropriate to deal with such imperfect data, two solutions are commonly considered: either ignoring such data by regarding them as unknown or incomplete, or developing suitable tools for treating them. In the first approach information is lost and this may lead to inaccurate classification models. On the contrary, if we adjust classification techniques in order to be able to deal with imperfect data, the models produced will describe the concepts more faithfully.

In this section, we extend the possibilistic classifiers previously presented in Section 4, in order to handle uncertainty in data representation. Before proposing solutions in order to deal with uncertainty in both the training and the testing data sets, we state the hypotheses we use for the representation of the uncertainty in the data sets.

5.1. Representation of uncertain training and testing data sets

The uncertain possibilistic classifier proposed in this section, is based on the following hypothesis:

- All training instances are assumed to have perfect (certain and precise) attribute values (as in the settings of possibilistic classifiers presented previously).
- All testing instances have imprecise attribute values modeled by intervals.
- The class of any training instance is represented through a possibility distribution over the class values thus reflecting uncertainty on the classification.

Let us for instance consider a classification problem with 3 class labels (c_1, c_2 and c_3). In case of a standard training instance a with M certain and precise numerical attributes, a unique class c_1 is assigned, e.g.:

$$a = (a_1, a_2, ..., a_M, c_1)$$

In the uncertainty version of instance a_U, the knowledge about the class associated with this example is represented by means of a possibility distribution over the different possible class labels, and then we have:

$$a_U = (a_1, a_2, ..., a_M, \pi_{c_1}, \pi_{c_2}, \pi_{c_3})$$

where π_{c_i} is the possibility degree for the class c_i. For instance, the distribution $(1, 0.3, 0.7)$ expresses that the expert finds the instance fully compatible with the first class, less compatible with the third one and still less compatible with the second one. There are some other noticeable particular cases that can also be represented as such an uncertainly classified instance. Thus the distribution $(1, 1, 0)$ represents pure imprecision over the first and the second class labels which are fully plausible whereas the third class is impossible. Besides, a distribution such as $(1, 0, 0)$ coincides with a standard, certainly classified instance for which only the first class is possible while the others are completely rejected. Finally, the expert may also express his total ignorance about the instance by choosing the distribution (1, 1, 1) according to which all class labels are fully plausible.

Uncertainty in the testing set concerns attribute values and each instance may include precise or imprecise attribute values. Since we are only interested in numerical data in this framework, the proposed model allows an expert to express his imprecise knowledge by means of an interval restricting the attribute value. Thus, for each imprecise attribute, the observed value is supposed to be the form of $I_i = [L_i, U_i]$ where L_i and U_i are respectively the lower and the upper bounds for the true attribute value a_i such that $L_i < a_i < U_i$. For imprecise attribute values, the degree of ignorance about the real value is related to the relative *width* of the interval for this attribute.

5.2. Processing of uncertain classes in the training set

Let Tr be a training set composed of N instances involving M numerical attributes. Instead of an exact class label, for each instance we assign a possibility distribution on the different possible labels. Our problem is to estimate a possibility distribution for each attribute a_i given the class c_j which can be the most specific representation for *uncertain* numerical data. The conditional possibility $\pi(a_i|c_j)$ is obtained by computing (using equation 7) the possibility distribution $\pi_{(N_j, \overline{X}_{(i,j)}, S_{(i,j)})}$ that encodes the confidence region $\mathcal{R}_{(i,j)}(N_j, \overline{X}_{(i,j)}, S_{(i,j)})$ for the parameters of the normal distributions of A_i. The class of an example being now pervaded with uncertainty, we use weighted sums for evaluating the values N_j, $\overline{X}_{(i,j)}$ and $S_{(i,j)}$. Then we have:

$$N_j = \sum_{k=1}^{N} \pi_k(c_j),$$

$$\overline{X}_{(i,j)} = \frac{\sum_{k=1}^{N} \pi_k(c_j) * a_{ik}}{N_j}$$

and

$$S_{(i,j)} = \frac{\sum_{k=1}^{N} \pi_k(c_j) * (a_{ik} - \overline{X}_{(i,j)})^2}{N_j}$$

where $\pi_k(c_j)$ is the possibility that example k belongs to class c_j. The choice a weighted sum for N_j corresponds to the standard definition of the scalar

cardinality of a fuzzy set [47], which leads to the other weighted sums in the same spirit. Besides, it is noticeable that counterpart of these three expressions can be found in the work of Côme et al [15] where the more general setting of belief functions (that encompasses possibility theory) is used for describing uncertain classes.

We note that the proposed model, supporting uncertainty in the class labels, also includes the certain case where $\pi_k(c_j)$ is 1 for the the true label and 0 otherwise.

For the FNPC-2, we extend equation (8) as follows:

$$\pi(a|c_j) = \frac{1}{N_j} \sum_{k=1}^{N_j} \pi_{(N_j,\mu_{ik},\sigma)}(a) * \pi_k(c_j) \tag{9}$$

with N_j computed as above and $\sigma = \frac{1}{\sqrt{N}}$.

5.3. Processing of imprecise attributes in the testing set

In the following we propose an algorithm for handling imprecision in attribute values in the testing set. Let us consider a function F that estimates conditional possibilities for attribute values in the perfect case. For each *observed* attribute value x_i, this function estimates $\pi_{(a_i|c_j)}(x_i)$. When the observed value of an attribute is no longer a fixed value in the domain of the attribute, but rather an interval I_i, the problem amounts to estimate $\pi_{(a_i|c_j)}(I_i)$.

In agreement with the definition of the possibility measure of the event I_i, Equation(9) is extended as follows :

$$\pi(I_i|c_j) = sup\{\pi(a_i|c_j), a_i \in I_i\} \tag{10}$$

This means that $\pi(I_i|c_j)$ is estimated as follows :

1. Search for all attribute values a_i in the training set such that $a_i \in I_i$
2. Compute the possibility of attribute values a_i given the class c_j by equation (9)
3. Consider the highest possibility as the possibility of I_i.

6. Experiments and discussion

This section provides experimental results of possibilistic classifiers for perfect and imperfect numerical data. The experimental study is based on several datasets taken from the U.C.I. repository of machine learning databases [49]. A brief description of these datasets is given in Table 1. Since we have chosen to deal only with numerical attributes in this study, all these datasets have numerical attribute values. For each dataset, we used a ten-fold cross validation to evaluate the generalization accuracy of classifiers.

The experimental study is divided in two parts. First, we evaluate the NPC-2 and FNPC-2 methods and compare our results to those of a classical NBC [42] and FNBC [42]. Second, we test the efficiency of the proposed possibilistic classifiers to support uncertainty related to the classes or attributes.

Table 1: Description of datasets

Database	**Data**	**Attributes**	**Classes**
Iris	150	4	3
W. B. Cancer	699	8	2
Wine	178	13	3
Diabetes	768	7	2
Magic gamma telescope	1074	10	2
Transfusion	748	4	2
Satellite Image	1090	37	6
Segment	1500	20	7
Yeast	1484	9	10
Ecoli	336	8	8
Glass	214	10	7
Iosophere	351	35	2
Block	5473	10	5
German	1000	25	2
Heart	270	14	2

6.1. Experiments of possibilistic classifiers for the perfect numerical data

This section provides experimental results with classical (non-uncertain) databases for possibilistic classifiers that have been previously introduced. In order to evaluate the accuracy of each classifier, we have used the standard *Percent of Correct Classification* (PCC) defined as follows:

$$Pcc = \frac{number\ of\ well\ classified\ instances}{total\ number\ of\ classified\ instances} * 100 \tag{11}$$

Table 2 shows the classification performance obtained with NPC-2, NBC, FNPC-2 and FNBC for the fifteen mentioned datasets. Let us point out that:
• A normality test (test of Shapiro-Wilk) performed on these databases shows that they contain attributes that are not normally distributed [10]. The NBC seems to partially fail when classifying instances in datasets with attributes that strongly violate the normality assumption.
• We expect that, if the normality assumption is strongly confirmed for a given dataset, it is better to use a probability distribution for classification since it remains more precise. In the other case, we may suppose that applying a probability-possibility transformation on the NBC (which leads to NPC [10] or to NPC-2) enables the classifier to be less sensitive to normality violation. As suggested in Section 2, one may also think that when normality assumption is not supported by the data, especially for datasets with a high number of attributes, the NBC reinforces the error rate (by the use of multiplication), making the NPC-2 more efficient in this case.
• As previously observed in [42], the FNBC is overall better than classical NBC. In fact, FNBC is more accurate than the NBC in 9 of the 15 datasets and

Table 2: Experimental results for perfect data given as the mean and the standard deviation of 10 cross-validations

	NPC-2	**NBC**	**FNPC-2**	**FNBC**	**SVM**
Iris	95.33±4.27	95.33±5.21	95.33±5.21	95.33±5.54	*96*
Cancer	95.46±2.02	96.19±0.97	**97.66±0.72**	**97.66±0.72**	*97.07*
Wine	95.48±4.18	97.15±2.86	**97.78±2.72**	96.6±3.73	*98.31*
Diabetes	72.91±5.51	75.52±2.67	**76.17±3.58**	75.64±3.56	*77.34*
Magic	59.32±6.33	65.93±2.91	**73.0±2.49**	72.26±2.39	*76.90*
Transfusion	74.36±6.32	75.02±5.56	**76.76±5.73**	75.7±6.19	*76.20*
Sat. Image	90.46±3.98	90.83±3.58	**91.28±3.16**	90.55±3.15	*94.22*
Segment	74.46±3.44	80.87±2.37	**91.13±2.73**	88.6±3.48	*91.93*
Yeast	57.68±3.36	46.97±4.69	**58.36±2.14**	52.9±3.73	*57.07*
Ecoli	83.08±5.47	81.27±5.16	**85.8±5.6**	75.82±7.1	*84.22*
Glass	46.32±14.79	43.12±8.12	**67.38±9.86**	57.97±8.98	*57.47*
Iosophere	60.95±9.1	70.09±6.15	**92.62±5.05**	92.05±5.0	*88.60*
Block	88.49±1.86	89.66±3.22	**93.51±1.07**	90.21±2.14	*92.92*
German	73.2±2.99	73.0±2.97	**75.7±2.93**	70.0±4.22	*76.40*
Heart	**84.45±4.32**	83.34±5.56	83.7±5.79	84.08±6.15	*84.07*

less accurate in 3 datasets and not significantly different in three cases ("Iris", "Diabetes" and "Satellite Image").

• For the four classifiers (NPC-2, NBC, FNPC-2 and FNBC), classification results of the FNPC-2 are better than the three other classifiers for all datasets except in the case of "Iris", "Cancer" and "Heart" databases where FNPC-2 has almost the same accuracy as others.

• If we compare results for the two flexible classifiers (FNPC-2 and FNBC), we note that the FNPC-2 performs better for a majority of datasets. For this classifier, the greatest increase in accuracy compared to the FNBC occurs for databases "Yeast", "Glass", "Ecoli", "Segment", "German" and "Block" (Table 2). In Table 1, we note that the number of attributes for these databases ranges from 8 to 25, and the number of classes from 5 to 10, except the "German" which has only 2 classes. The FNPC-2 classifier is significantly more efficient than FNBC one (and also than NPC-2 and NBC) for datasets with a high number of attributes and classes.

• To compare the four classifiers in terms of PCC, we use the Wilcoxon Matched-Pairs Signed-Ranks Test as proposed by Demsar [18]. It is a non-parametric alternative to the paired t-test that enables us to compare two classifiers over multiple data sets. Comparison results given in Table 3 show that the FNPC-2 is always significantly better ($p - value < 0.05$) than the three other classifiers for all data sets whereas the two naïve classifiers (NPC-2 and NBC) have competitive performance. These results confirm those reported in [11].

In Table 2, we also report classification accuracy of the SVM classifier for

the fifteen datasets. The reported results are obtained by applying the WEKA software [56] implementation of the support vector classifier (the SMO class).

If we compare results for the FNPC-2 and the SVM classifiers, we note that the FNPC-2 is more accurate than the SVM in 7 datasets and less accurate in the remaining. In particular, the SVM is significantly better in the dataset "Diabetes", "Magic" and "Sat. Image". However the FNPC-2 is significantly more efficient for the datasets "Glass", "Ecoli", "Iosophere", "Yeast" and "Block". As we have already noticed, possibilistic classifiers seem to perform well for datasets with large dimension (if compared to Bayesian classifiers or SVM).

Table 3: Results for the Wilcoxon Matched-Pairs Signed-Ranks Test

FNPC-2 Vs NPC-2	**FNPC-2 Vs NBC**	**FNPC-2 Vs FNBC**	**NPC-2 Vs NBC**
$p \leq 0.001$	$p \leq 0.0006$	$p \leq 0.0035$	$p \leq 0.376$

6.2. Experiments of possibilistic classifiers for the imperfect numerical data

Even if uncertainty in databases may be regarded as an important issue in machine learning, there are no uncertain nor imprecise data sets which could be used for testing algorithms dealing with such type of data. For this reason, we first give here a heuristics to introduce uncertainty and imprecision in an artificial manner. In the second part of this section we present the criteria suitable for evaluating the classification accuracy of the FNPC-2 in the imperfect case. Finally, we give results for this case.

6.2.1. Generation of imperfect data

Data sets described in Table 1 are initially perfect with certain and precise attributes and classes. In order to evaluate the FNPC-2 in the imperfect case, we have artificially introduced imperfection in these data sets by transforming the original precise and certain instances into imperfect ones.

Creating possibilistic labels: Uncertainty on the training set is created by replacing the certain class label of each instance by a possibility distribution over class labels. To generate a possibility distribution, we are going to simulate the fact that we have two experts and that they are, to some extent, unable to classify each training instance in a certain manner. So, each expert gives a possibility distribution over class labels reflecting his knowledge about this uncertain situation. Then we apply an information fusion procedure [28] to produce the final possibility distribution for each instance. In practice, each expert will simply be a possibilistic classifier trained on the perfect (certain and precise) data set. In this experiment we have used the FNPC and FuHC classifiers [10] to simulate two experts. For information fusion, we apply a *disjunctive operator* [28] to create the final possibility distribution $\pi_{a_{tr}}$:

$$\forall \omega \in \Omega, \pi_{\vee}(\omega) = \oplus_{i=1,\ldots,n} \pi_i(\omega) = \max_{i=1}^{n} \pi_i(\omega) \quad (12)$$

We prefer to use here the disjunctive operator to the conjunctive one since the two classifiers may disagree and we cannot be sure which one is the more reliable. Moreover, possibilistic distributions generated with this operator cover the imprecise case where more than one class may have a possibility degree equal to 1. We create uncertain training set in the following way:

1. Train the FNPC and the FuHC using the original crisp training set.
2. Use the obtained possibilistic classifiers to predict the class labels for each training instance.
3. For each training instance a_{tr}, use the two possibility distributions obtained from each classifier using a *disjunctive operator*.
4. Keep the attribute values of each instance in the training set unchanged and replace the crisp class label by $\pi_{a_{tr}}$.

Creating imprecise attributes values: Attributes in the testing set are made imprecise in the following way. In each testing instance, we convert each attribute value into an uncertain interval. We first compute the range for each attribute ($[X_{min}, X_{max}]$). Then we replace each attribute value x by a generated interval I = [L, U] in order to create imprecision on this attribute. Lower bound L (resp. upper bound U) is calculated as follows: $L = x - (x - X_{min}) * rand1$ (resp. $U = x + (X_{max} - x) * rand2$), where $rand1$ (resp. $rand2$) denotes a random number bounded by $AttLev$. $AttLev$ is a level that refers to the width of the interval and takes values in $\{0.25, 0.5, 0.75$ or $1\}$. For each level $AttLev$, we generate an uncertain dataset U_{AttLev} where $rand1$ and $rand2$ range between 0 and $AttLev$. Hence, for each perfect testing set, we create four uncertain datasets $U_{0.25}, U_{0.5}, U_{0.75}$ and U_1.

6.2.2. Classification accuracy measures

To measure the accuracy of the FNPC-2, we use two evaluation criteria:

- **The percentage of Most Plausible Correct Classification (MPcc):** counts the percentage of instances whose all most plausible classes, predicted by the possibilistic classifier, are *exactly the same* as their initial most plausible classes given by the possibility distribution labeling each testing instance.

$$MPcc = \frac{Number\_of\_exactly\_well\_classified\_instances}{Total\_Number\_classified\_instances} * 100 \quad (13)$$

- **The Information Affinity-based Criterion: AffC** [41] is a degree of affinity between the predicted and the real possibility distribution labeling the testing instances which ranges in [0,1]:

$$InfoAffC = \frac{\sum_{i=1}^{n} Aff(\pi_i^{real}, \pi_i^{pred})}{Total\_Number\_classified\_instances} \quad (14)$$

$$Aff(\pi_1, \pi_2) = 1 - \frac{d(\pi_1, \pi_2) + Inc(\pi_1, \pi_2)}{2} \quad (15)$$

where $d(\pi_1, \pi_2)$ is the Manhattan distance between π_1 and π_2 and $Inc(\pi_1, \pi_2) = Inc(\pi_1 \wedge \pi_2)$ is the degree of inconsistency between π_1 and π_2 calculated as follows:

$$Inc(\pi) = 1 - \max_{\omega \in \Omega} \{\pi(\omega)\}$$

6.2.3. Results for the imperfect numerical data

This experimental study is divided in two parts. First, we evaluate the uncertain FNPC-2 to handle uncertainty only in class attribute and we keep attributes in the untouched testing set. Second, we test the accuracy of the proposed classifier when attributes in the testing set are uncertain whereas training set is kept untouched. We choose to test each uncertainty type independently in order to check the efficiency of the FNPC-2 to deal with each situation.

6.2.4. Uncertainty type 1: Uncertain classes

Table 4: Experimental results for uncertain classes given as the mean and the standard deviation of 10 cross-validations

	FNPC-2	
	MPcc	**AffC**
Iris	94.0±3.6	0.94±0.01
Cancer	96.19±1.8	0.98±0.0
Wine	92.64±6.2	0.94±0.01
Diabetes	76.85±6.4	0.96±0.0
Magic	74.4±3.7	0.93±0.0
Transfusion	83.57±4.2	0.98±0.0
Sat. Image	90.55±3.1	0.98±0.01
Segment	70.93±5.1	0.92±0.01
Yeast	58.28±3.6	0.96±0.0
Ecoli	80.36±8.2	0.93±0.01
Glass	52.54±13.3	0.92±0.02
Iosophere	78.63±6.9	0.94±0.02
Block	78.69± 1.07	0.94±0.0
German	81.2±3.9	0.96±0.01
Heart	89.26±6.5	0.96±0.01

Table 4 shows the classification performance (MPcc and InfoAffC criterion) obtained with the FNPC-2 for the 15 uncertain data sets.

If we analyze results in Table 4, we note that:

• As reported in the perfect case from these results we can say that, overall the FNPC-2 shows a high ability to detect the most plausible classes even for

uncertain datasets with high level of uncertainty (all training instances are uncertain).
• By analyzing the InfoAffC criteria we can see that the values are very high for the uncertain FNPC-2 and for all data sets (the InfoAffC is > 0.9). From these results, we can conclude that the proposed classifier is able to predict faithful possibility distributions.
• For the majority of data sets, the InfoAffC criteria confirms the results reported by the MPcc. However we can see a significant divergence between the values of InfoAffC and MPcc for some data sets (for example for the Segment, Glass, Iosophere and Block there is a significant decrease in MPcc, if compared to the perfect case which is respectively about 20 %, 15%, 13% and 15% however the InfoAffC remains higher than 0.9). This divergence means that for many testing instances, the FNPC-2 provides possibility degrees *close* to the initial possibility distribution (high values of InfoAffC) but the predicted and real full plausible classes are *not exactly* the same (low values of MPcc). So we can say that this decrease in accuracy for these datasets due to the rigid nature of the MPcc criteria which causes the absence of classification for many instances in the data set where the classifier provides more than one fully plausible class which are not exactly the same as those given in the real distribution. This mainly happens for datasets having a large number of classes.

6.2.5. Uncertainty type 2: Imprecise attributes

Table 5 shows the MPcc and the InfoAffC results obtained with the FNPC-2 for each imprecision level on attributes and for the fifteen mentioned data sets. By comparing the classification performance on each imprecision level we see that accuracy decreases when the imprecision level of attributes increases (when intervals become broader). Despite this decrease we note that:
• As in the uncertainty type 1 case (Table 4), the FNPC-2 has reported relatively good performance if compared to the perfect case. We can also remark that the decrease of accuracy is relatively smooth.
• Despite the decrease in accuracy, we note it remains acceptable in average. For instance, if we analyze results in Table 5, we remark that the MPcc remains higher than 60% for the highest uncertainty level (U_1)(the worst case) and this for all data sets except the "Yeast" and "Glass" where the values are respectively about 31% and 43%. The low results reported for these two data sets are not specifically due to to the FNPC-2 since the MPcc reported for the original version for the certain case of these data sets is only about 58% for the Yeast and 67% for the Glass for FNPC-2.
• The values of the InfoAffC criterion reported for the FNPC-2 and for the different data sets are relatively good. For 9 of fifteen datasets, this value remains higher than 0.8 (for all uncertainty levels) and it is higher than 0.7 for the remaining data sets. Thus, we can conclude that the predicted and initial possibility distributions are relatively consistent.

From results given in Tables 4 and 5, FNPC-2 appears to be accurate and can be considered as a competitive classifier which is well suited for dealing with perfect or imperfect continuous data.

Table 5: Experimental results for uncertain attributes given as the mean and the standard deviation of 10 cross-validations

	$U_{0.25}$		$U_{0.5}$		$U_{0.75}$		U_1	
	MPcc	**AffC**	**MPcc**	**AffC**	**MPcc**	**AffC**	**MPcc**	**AffC**
Iris	93.33±8.9	0.95±0.05	90.67±10.0	0.94±0.06	88.0±11.9	0.92±0.05	86.67±9.4	0.9±0.04
Cancer	97.66±1.5	0.98±0.02	96.04±2.5	0.96±0.02	95.9±2.3	0.96±0.02	94.28±2.6	0.94±0.02
Wine	95.48±3.4	0.97±0.02	94.93±3.9	0.96±0.03	91.6±3.7	0.94±0.02	85.9±5.3	0.9±0.04
Diabetes	73.17±3.6	0.77±0.02	69.52±4.5	0.75±0.03	68.35±4.8	0.74±0.04	66.0±4.2	0.72±0.02
Magic	73.29±4.2	0.78±0.02	72.46±3.2	0.78±0.02	72.26±3.1	0.77±0.02	70.77±3.9	0.76±0.02
Transfusion	65.41±7.9	0.73±0.04	63.82±7.1	0.73±0.04	63.02±7.5	0.72±0.03	60.86±5.4	0.72±0.03
Sat. Image	89.27±2.1	0.93±0.01	86.15±3.2	0.91±0.02	85.78±2.6	0.91±0.02	84.4±3.5	0.9±0.02
Segment	87.93±2.3	0.93±0.01	83.13±3.6	0.91±0.02	77.8±3.2	0.89±0.02	73.4±3.6	0.86±0.02
Yeast	53.71±2.8	0.79±0.02	49.33±4.7	0.77±0.01	40.51±3.1	0.74±0.02	31.2±3.0	0.7±0.01
Ecoli	81.27±5.3	0.91±0.03	77.19±9.9	0.9±0.04	70.28±10.5	0.86±0.03	63.65±6.5	0.83±0.02
Glass	49.56±12.57	0.78±0.03	47.43±15.5	0.78±0.04	45.34±12.05	0.76±0.04	43.62±14.33	0.76±0.05
Iosophere	91.16±2.0	0.92±0.02	90.3±3.4	0.91±0.03	89.74±4.3	0.91±0.04	86.9±5.4	0.88±0.04
Block	89.2±1.7	0.93±0.01	86.86±2.2	0.91±0.01	81.2±2.0	0.87±0.01	75.9±3.0	0.84±0.01
German	71.5±4.2	0.76±0.02	71.8±4.3	0.76±0.02	69.4±3.2	0.75±0.02	69.2±3.4	0.74±0.03
Heart	84.08±4.7	0.85±0.04	81.85±5.8	0.84±0.04	81.85±6.5	0.84±0.05	81.48±7.0	0.84±0.05

7. Conclusion

Imperfection in databases, including imprecision and uncertainty, is gaining more attention since decision system may have to deal with such a kind of information. Most of possibilistic classifiers [37] [40] that have been proposed with this concern are only suitable for discrete attributes. This work has investigated a possibilistic classification paradigm that may be viewed as a counterpart of Bayesian classification and that applies to continuous attribute domains. Then an important issue is the estimation of possibilistic distributions from numerical data, without discretization. For this purpose, we have proposed and tested the performance of two possibilistic classifiers which are variants of those previously proposed in [10] and [12], called the NPC-2 and the FNPC-2.

For these classifiers, we have used a probability-possibility transformation method enabling us to derive a possibilistic distribution as a family of Gaussian distributions. First, we have applied the transformation method to move from a classical probabilistic NBC to NPC-2, which takes into account the confidence intervals of the Gaussian distributions by considering the amount of data available for estimating the parameters of the distributions. Then, we have tested the feasibility of a Flexible Naive Possibilistic Classifier (FNPC-2), which is the possibilistic counterpart of the Flexible Naive Bayesian Classifier. The FNPC-2 estimates possibilistic distributions in a non-parametric way by applying the transformation method to kernel densities instead of Gaussian ones. The rationale behind this classifier is that kernel densities are less sensible than Gaussian

ones to normality violation.

The second interest of this paper is to extend the proposed possibilistic classifier for handling uncertainty in data sets. Two types of uncertainty are considered: i) uncertainty related to class attribute in the training set modeled through possibility distributions over class labels, and ii) uncertainty related to attribute values in the testing set represented by intervals for continuous data. For the first type of uncertainty, we have adapted the possibilistic classification model suitable for the certain case, to support uncertainty in class labels. We have also proposed an algorithm based on the extension principle to deal with the imprecision of attribute values. The algorithm estimates possibility distributions for an interval-valued attribute by looking for the possibility distributions associated with each attribute value in the training set belonging to this interval.

To test possibilistic classifiers in the uncertain case, we have artificially introduced imperfection in data sets from the UCI machine learning repository [49]. Experimental results show the performance of these classifiers for handling numerical input data. However, while the NPC-2 is less sensible than NBC to normality violation, the FNPC-2 shows high classification accuracy and good ability to deal with any type of data when compared to the NPC-2 and to Bayesian classifiers. Regarding accuracy, we also show that FNPC-2 compete with SVM in the perfect information case. Results for the imperfect data shows the efficiency of the FNPC-2 to predict the class labels from possibility distributions that are quite consistent with initial distributions.

As future work, it would be interesting to deal with uncertain attribute values both in the training and testing sets. Estimating conditional distributions from training set including uncertain attributes and classes at the same time is a more tricky issue to be considered. In addition, in order to really exploit the proposed possibilistic classifiers for uncertain data, it would be important to test possibilistic approaches on genuine uncertain data.

References

[1] A. Aregui and T. Denoeux. Consonant belief function induced by a confidence set of pignistic probabilities. In K. Mellouli, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 4724 of *LNCS*, pages 344–355. Springer Berlin / Heidelberg, 2007.

[2] B. C. Arnold and R. M. Shavelle. Joint confidence sets for the mean and variance of a normal distribution. *The American Statistician*, 52(2):133–140.

[3] N. Ben Amor, S. Benferhat, and Z. Elouedi. Qualitative classification and evaluation in possibilistic decision trees. In *FUZZ-IEEE'04*, volume 1, pages 653–657, 2004.

[4] N. Ben Amor, K. Mellouli, S. Benferhat, D. Dubois, and H. Prade. A theoretical framework for possibilistic independence in a weakly ordered

setting . *Inter. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:117–155, 2002.

[5] S. Benferhat and K. Tabia. An efficient algorithm for naive possibilistic classifiers with uncertain inputs. In *Proc. 2nd Inter. Conf. on Scalable Uncertainty Management (SUM'08), Springer LNAI*, volume 5291, pages 63–77, 2008.

[6] J. Beringer and E. Hüllermeier. Case-based learning in a bipolar possibilistic framework. *Inter. Journal of Intelligent Systems*, 23:1119–1134, 2008.

[7] C. M. Bishop. Latent variable models. In *Learning in Graphical Models*, pages 371–403. MIT Press, 1999.

[8] C. Borgelt and J. Gebhardt. A naïve Bayes style possibilistic classifier. In *Proc. 7th European Congress on Intelligent Techniques and Soft Computing*, pages 556–565, 1999.

[9] C. Borgelt and R. Kruse. Efficient maximum projection of database-induced multivariate possibility distributions. In *Proc. 7th IEEE Inter. Conf. on Fuzzy Systems*, pages 663–668, 1998.

[10] M. Bounhas, K. Mellouli, H. Prade, and M. Serrurier. From Bayesian classifiers to possibilistic classifiers for numerical data. In Deshplande A. and Hunter A., editors, *Proc. of The Fourth Inter. Conf. on Scalable Uncertainty Management (SUM10)*, volume LNAI 6379, pages 112–125. Springer-Verlag, 2010.

[11] M. Bounhas, K. Mellouli, H. Prade, and M. Serrurier. Possibilistic classifiers for numerical data. *Soft Computing*, 17:733–751, May 2013.

[12] M. Bounhas, H. Prade, M. Serrurier, and K. Mellouli. Possibilistic classifiers for uncertain numerical data. In *Proc. of the 11th European conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2011)*, volume LNAI 6717, pages 434–446. Springer-Verlag, 2011.

[13] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, pages 101–107, 1999.

[14] M. R. Civanlar and H. J. Trussell. Constructing membership functions using statistical data. *Fuzzy Sets and Systems*, 18:1–13, January 1986.

[15] E. Côme, L. Oukhellou, T. Denoeux, and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334 – 348, 2009.

[16] G. De Cooman. Possibility theory. Part I: Measure- and integral-theoretic ground- work; Part II: Conditional possibility; Part III: Possibilistic independence. *Inter. Journal of General Systems*, 25:291–371, 1997.

[17] M. Delgado. On the concept of possibility-probability consistency. *Fuzzy Sets and Systems*, 21(3):311–318, 1987.

[18] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[19] A. Denton and W. Perrizo. A kernel-based semi-naïve Bayesian classifier using p-trees. In *Proc. of the 4th SIAM Inter. Conf. on Data Mining*, 2004.

[20] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning*, 29:102–130, 2002.

[21] D. Dubois. Possibility theory and statistical reasoning. *Computational Statistics and Data Analysis*, 51:47–69, 2006.

[22] D. Dubois, L. Foulloy, G. Mauris, and H. Prade. Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities. *Reliable Computing*, 10:273–297, 2004.

[23] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[24] D. Dubois and H. Prade. The logical view of conditioning and its application to possibility and evidence theories. *Inter. Journal of Approximate Reasoning*, 4:23–46, 1990.

[25] D. Dubois and H. Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

[26] D. Dubois and H. Prade. Fuzzy sets and probability : Misunderstandings, bridges and gaps. In *Proc. of the Second IEEE Conf. on Fuzzy Systems*, pages 1059–1068. IEEE, 1993.

[27] D. Dubois and H. Prade. Possibility theory: Qualitative and quantitative aspects. *D. Gabbay and P. Smets. editors. Handbook on Defeasible Reasoning and Uncertainty Management Systems*, 1:169–226, 1998a.

[28] D. Dubois and H. Prade. An overview of ordinal and numerical approaches to causal diagnostic problem solving. *D.M. Gabbay, R. Kruse editors. Abductive Reasoning and Learning, Handbooks of Defeasible Reasoning and Uncertainty Management Systems*, 4:231–280, 2000.

[29] D. Dubois and H. Prade. Formal representations of uncertainty. In *Decision-making - Concepts and Methods. D. Bouyssou, D. Dubois, M. Pirlot, H. Prade Editors*, chapter 3, pages 85–156. Wiley, 2009.

[30] D. Dubois, H. Prade, and S. Sandri. On Possibility/Probability Transformations. In R. Lowen and M. Roubens, editors, *Fuzzy Logic*, volume 12, pages 103–112. Springer Netherlands, 1993.

[31] M.A.T. Figueiredo, J.M.N. Leitãlo, and A.K. Jain. On fitting mixture models. In E. R. Hancock and M. Pelillo, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 1654 of *LNCS*, pages 54–69. Springer Berlin Heidelberg, 1999.

[32] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–161, 1997.

[33] D. Geiger and D. Heckerman. Learning gaussian networks. *Technical report, Microsoft Research, Advanced Technology Division*, 1994.

[34] D. Grossman and P. Dominigos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proc. Inter. Conf. of Machine Learning*, pages 361–368. ACM Press, 2004.

[35] M. Ghasemi Hamed, M. Serrurier, and N. Durand. Possibilistic kNN regression using tolerance intervals. In *IPMU 2012, Catania, Italy, July*, volume 299 of *Communications in Computer and Information Science*. Springer, 2012.

[36] M. Ghasemi Hamed, M. Serrurier, and N. Durand. Representing uncertainty by possibility distributions encoding confidence bands, tolerance and prediction intervals. In *Proc. 6th Inter. Conf. on Scalable Uncertainty Management (SUM'12)*, volume 7520 of *LNCS*, pages 233–246. Springer, 2012.

[37] B. Haouari, N. Ben Amor, Z. Elouadi, and K. Mellouli. Naive possibilistic network classifiers. *Fuzzy Set and Systems*, 160(22):3224–3238, 2009.

[38] E. Hüllermeier. Possibilistic instance-based learning. *Artificial Intelligence*, 148(1-2):335–383, 2003.

[39] E. Hüllermeier. Fuzzy methods in machine learning and data mining : Status and prospects. *Fuzzy Sets and Systems*, 156(3):387–406, 2005.

[40] I. Jenhani, N. Ben Amor, and Z. Elouedi. Decision trees as possibilistic classifiers. *Inter. Journal of Approximate Reasoning*, 48(3):784–807, 2008.

[41] I. Jenhani, S. Benferhat, and Z. Elouedi. Learning and evaluating possibilistic decision trees using information affinity. In *Inter. Journal of Computer Systems Science and Engineering*, volume 4(3), pages 206–212, 2010.

[42] G. H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *proc. of the 11th Conf. on Uncertainty in Artificial Intelligence*, 1995.

[43] G. J. Klir. A principle of uncertainty and information invariance. *Inter. Journal of General Systems*, 17(23):249–275, 1990.

[44] I. Kononenko. Semi-naive Bayesian classifier. In *Proc. of the European Working Session on Machine Learning*, pages 206–219, 1991.

[45] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proc. of AAAI-92*, volume 7, pages 223–228, 1992.

[46] P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proc. of 10th Conf. on Uncertainty in Artificial Intelligence UAI-94*, pages 399–406, 1994.

[47] A. De Luca and S. Termini. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, 20(4):301 – 312, 1972.

[48] G. J. McLachlan and D. Peel. Finite mixture models. *Probability and Mathematical Statistics, John Wiley and Sons*, 2000.

[49] J. Mertz and P. M. Murphy. UCI repository of machine learning databases. *Available at: ftp://ftp.ics.uci.edu/pub/machine-learning-databases*.

[50] A. Pérez, P. Larraoaga, and I. Inza. Bayesian classifiers based on kernel density estimation: Flexible classifiers. *Inter. Journal of Approximate Reasoning*, 50:341–362, 2009.

[51] B. Qin, Y. Xia, and F. Li. A Bayesian classifier for uncertain data. *In the 25th ACM Symposium on Applied Computing (SAC)*, pages 1010–1014, 2010.

[52] B. Qin, Y. Xia, S. Prabhakar, and Y. Tu. A rule-based classification algorithm for uncertain data. In *IEEE Inter. Conf. on Data Engineering*, pages 1633–1640, 2009.

[53] M. Sahami. Learning limited dependence Bayesian classifiers. In *Proc. of the 2nd Inter. Conf. on Knowledge Discovery and Data Mining*, pages 335–338, 1996.

[54] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[55] T. Denœux T. and L.M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Set and Systems*, 122(3):47–62, 2001.

[56] H.I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edition Morgan Kaufmann Publishers, 2005.

[57] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[58] H. Zhang. The optimality of naive Bayes. In *Proc. of 17th Inter. FLAIRS Conf. (FLAIRS2004)*. AAAI Press, 2004.

Appendix: Naive Bayesian Classifiers

Naive Bayesian Classifiers (NBC) are based on Bayes rule. They assume the independence of the input variables. Despite their simplicity, NBC can often outperform more sophisticated classification methods [45]. A NBC can be seen as a Bayesian network in which predictive attributes are assumed to be conditionally independent given the class attribute.

Given a vector $X = \{x_1, x_2, ..., x_n\}$ to be classified, a NBC computes the posterior probability $P(c_j|X)$ for each class c_j in a set of possible classes $C = (c_1, c_2, ..., c_m)$, and labels the case X with the class c_j that achieves the highest posterior probability, that is:

$$c^* = arg \max_{c_j} P(c_j|X) \tag{.1}$$

Using the Bayes rule:

$$P(c_j|x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n|c_j) * P(c_j)}{P(x_1, x_2, ..., x_n)} \tag{.2}$$

The denominator $P(x_1, x_2, ..., x_n)$ is a normalizing factor that can be ignored when determining the maximum *posterior* probability of a class, as it does not depend on the class. The key term in equation (.2) is $P(x_1, x_2, ..., x_n|c_j)$ which is estimated from training data. Since Naive Bayes assumes that conditional probabilities of attributes are statistically independent we can decompose the likelihood into a product of terms:

$$P(x_1, x_2, ..., x_n|c_j) = \prod_{i=1}^{n} p(x_i|c_j) \tag{.3}$$

Even under the independence assumption, the NBC have shown good performance for datasets containing dependent attributes. Domingos and Pazzani [20] explain that attribute dependency does not strongly affect the classification accuracy. They also relate good performance of NBC to the zero-one loss function which considers that a classifier is successful when the maximum probability is assigned to the correct class (even if estimated probability is inaccurate). The work in [58] gives a deeper explanation about the reasons for which the efficiency of NBC is not affected by attribute dependency. The author shows that, even if attributes are strongly dependent (if we look at each pairs of attributes), the global dependencies among all attributes could be insignificant because dependencies may cancel each other out and so they do not affect classification.

The most well-known Bayesian classification approach uses an estimation based on a multinomial distribution over the discretized variables, and leads to so-called multinomial classifiers. Such a classifier, which handles only discrete attributes (continuous attributes must be discretized), assumes that all attributes follow a multinomial probability distribution. A variety of multinomial classifiers have been proposed for handling an arbitrary number of independent attributes. Let us mention especially [45], [46], [34], semi-naive Bayesian

classifiers [44], [19], tree-augmented naive Bayesian classifiers [32], k-dependence Bayesian classifiers [53], and Bayesian Network-augmented naive Bayesian classifiers [13].

A second family of NBC is suitable for continuous attribute values. They directly estimate the true density of attributes using *parametric* density. A supplementary common assumption made by the NBC in that case, is that within each class the values of numeric attributes are normally distributed around the mean, and they model each attribute through a single Gaussian distribution. Then, the NBC represent such a distribution in terms of its *mean* and *standard deviation* and compute the probability of an observed value from such estimates. This probability is calculated as follows:

$$p(x_i|c_j) = g(x_i, \mu_j, \sigma_j) = \frac{1}{\sqrt{2\Pi}\sigma_j} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}} \tag{.4}$$

The Gaussian classifiers [33] [42] are known for their simplicity and have a smaller complexity, compared to other non-parametric approximations. Although the normality assumption may be a valuable approximation for many benchmarks, it is not always the best estimation. Moreover, if the normality assumption is violated, classification results of NBC may deteriorate.

Other approaches using a non-parametric estimation are those breaking with the strong parametric assumption. The main approaches are based on the mixture model [31] [48] and the Gaussian mixture models [7] [48]. Other approaches use kernel densities [42][50], leading to so-called Flexible Classifiers. This name is due to the ability of such classifier to represent densities with more than one mode in contrast with simple Gaussian classifiers. Flexible classifiers represent densities of different shapes with high accuracy; however it results into a considerable increase in complexity.

John and Langley [42] have proposed a Flexible Naive Bayesian Classifier (FNBC) that abandons the normality assumption and instead uses nonparametric kernel density estimation for each conditional distribution. The FNBC has the same properties as those introduced for the NBC, the only difference is instead of estimating the density for each continuous attribute x by a single Gaussian $g(x, \mu_j, \sigma_j)$, this density is estimated using an averaged large set of Gaussian kernels. To compute continuous attribute density for a specific class j, FNBC calculates n Gaussian distributions, where each of them stores each attribute value encountered during training for this class and then takes the average of the n Gaussians in order to estimate $p(x_i|c_j)$. More formally, probability distribution is estimated as follows:

$$p(x_i|c_j) = \frac{1}{N_j} \sum_{k=1}^{N_j} g(x_i, \mu_{ik}, \sigma_j) \tag{.5}$$

where k ranges over the training set of attribute x_i in class c_j, N_j is the number of instances belonging to the class c_j. The mean μ_{ik} is equal to the real value

of attribute i of the instance k belonging to the class j, e.g. $\mu_{ik} = x_{ik}$. For each class j, FNBC estimates this standard deviation by:

$$\sigma_j = \frac{1}{\sqrt{N_j}} \tag{.6}$$

The authors also prove kernel estimation consistency using equation (.6), (see [42] for details). It has been shown that the kernel density estimation used in the FNBC and applied on several datasets, enables this classifier to perform well in datasets where the parametric assumption is violated with little cost for datasets where it holds.

Pérez et al. [50] have recently proposed a new approach for Flexible Bayesian classifiers based on kernel density estimation that extends the FNBC proposed by [42] in order to handle dependent attributes and abandons the independence assumption. In this work, three classifiers: tree-augmented naive Bays, a k-dependence Bayesian classifier and a complete graph are adapted to the support kernel Bayesian network paradigm.